



SNESTIK

Seminar Nasional Teknik Elektro, Sistem Informasi,
dan Teknik Informatika

<https://ejurnal.itats.ac.id/snestik> dan <https://snestik.itats.ac.id>



Informasi Pelaksanaan :

SNESTIK III - Surabaya, 11 Maret 2023

Ruang Seminar Gedung A, Kampus Institut Teknologi Adhi Tama Surabaya

Informasi Artikel:

DOI : 10.31284/p.snestik.2023.4156

Prosiding ISSN 2775-5126

Fakultas Teknik Elektro dan Teknologi Informasi-Institut Teknologi Adhi Tama Surabaya
Gedung A-ITATS, Jl. Arief Rachman Hakim 100 Surabaya 60117 Telp. (031) 5945043
Email : snestik@itats.ac.id

Penentuan Keakuratan Kelompok Data Gambar pada Proses Segmentasi Menggunakan Algoritma Random Forest

Salma N. Aini¹, Dian P. Hapsari², Aeri Rachman³

Institut Teknologi Adhi Tama Surabaya^{1,2}, Universitas Trunojoyo³

e-mail: salmanraini79@gmail.com

ABSTRACT

Abstract. Most traditional segmentation methods are based on pixel intensities and spatial relationships, or constrained models found through optimization. Nonetheless, humans use much more knowledge when performing manual segmentation. Therefore, in recent years, trainable machine learning methods have emerged as a powerful tool to include some of that knowledge in the segmentation process and improve the accuracy of labeled regions. In this paper, an analysis is conducted to see how accurate image segmentation is using the random forest Algorithm. In this paper, we will review the results of comparing the performance of the random forest algorithm with the J48, Naïve Bayes, and Logistic regression algorithms. The results of the comparison of the several algorithms Random Forest has the highest accuracy of 97.7%.

Keywords: artificial Intelligent, segmenting, random forest.

ABSTRAK

Abstrak. Sebagian besar metode segmentasi tradisional didasarkan pada intensitas dan hubungan spasial piksel, atau model terbatas yang ditemukan melalui pengoptimalan. Meskipun demikian, manusia menggunakan lebih banyak pengetahuan saat melakukan segmentasi manual. Oleh karena itu, dalam beberapa tahun terakhir, metode pembelajaran mesin yang dapat dilatih telah muncul sebagai alat yang ampuh untuk menyerahkan sebagian dari pengetahuan tersebut dalam proses segmentasi dan meningkatkan akurasi wilayah berlabel. Pada paper ini dilakukan analisis untuk melihat seberapa akurat segmentasi gambar dengan menggunakan algoritma random forest. Dalam makalah ini akan diulas tentang hasil

perbandingan kinerja algoritma random forest dengan algoritma J48, Naïve bayes, dan Logistic regression. Hasil perbandingan dari beberapa algoritma tersebut Random Forest memiliki keakuratan tertinggi 97.7%.

Kata Kunci: kecerdasan buatan, segmenting, random forest.

PENDAHULUAN

Manusia pandai melihat gambar dan menemukan pola atau membuat perbandingan. Namun bisakah manusia membandingkannya secara kuantitatif? manusia menggunakan teknologi pembelajaran mesin untuk mengekstraksi informasi bermakna dari gambar yang tidak bisa dilakukan manusia. Ilmuwan telah mengembangkan metode pembelajaran mesin untuk menganalisis dan membandingkan gambar secara kuantitatif. Algoritma dalam pembelajaran mesin memberikan informasi yang kaya dan mendetail. Kemampuan ini dapat mempercepat waktu penelitian bagi seorang ahli dan pada akhirnya dapat digunakan untuk mempercepat proses penemuan selanjutnya [1].

Terdapat peningkatan kebutuhan untuk solusi analisis gambar otomatis. Sebelum dilakukan kegiatan analisis, perlu untuk memahami struktur kepentingan yang harus dideteksi dan didefinisikan menurut representasi yang cocok untuk kuantifikasi oleh komputer. Hal ini dapat dicapai melalui kegiatan segmentasi. Segmentasi citra atau gambar bertujuan untuk menyederhanakan penggambaran citra ke dalam bentuk yang lebih bermakna dan lebih mudah dianalisis. Proses mempartisi gambar menjadi beberapa wilayah atau segmen yang homogen. Segmentasi merupakan transisi utama dalam analisis gambar, menggantikan nilai intensitas dengan label wilayah [2].

Sebagian besar metode segmentasi tradisional didasarkan pada intensitas dan hubungan spasial piksel, atau model terbatas yang ditemukan melalui pengoptimalan. Meskipun demikian, manusia menggunakan lebih banyak pengetahuan saat melakukan segmentasi manual. Oleh karena itu, dalam beberapa tahun terakhir, metode pembelajaran mesin yang dapat dilatih telah muncul sebagai alat yang ampuh untuk menyerahkan sebagian dari pengetahuan tersebut dalam proses segmentasi dan meningkatkan akurasi wilayah berlabel. Dalam penelitian ini akan dilakukan implementasi algoritma Random Forest untuk kegiatan segmentasi citra, serta dilakukan perbandingan akurasi dengan beberapa algoritma yang lain [3][4].

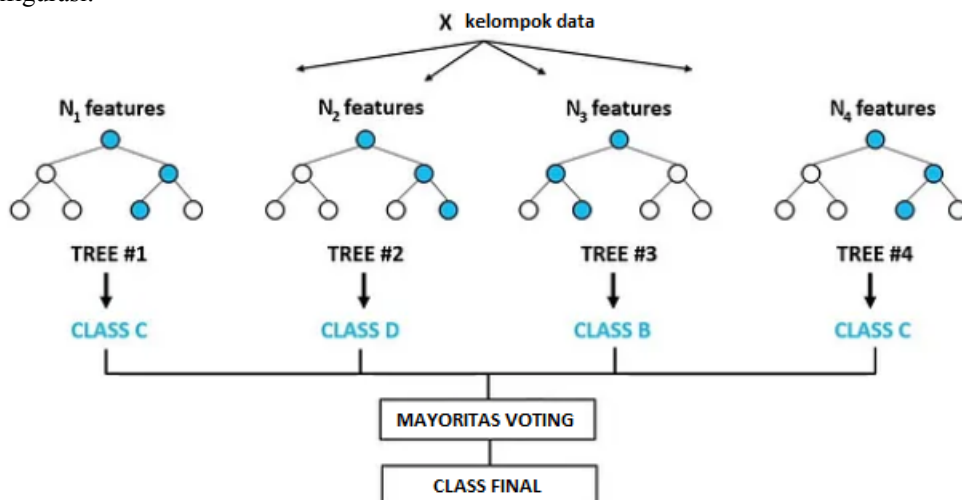
Dilakukan analisis perbandingan untuk pengklasifikasian gambar-gambar tersebut dengan menggunakan 4 algoritma seperti Naive Bayes, J48, Logistic Regression [5][6][7]. Pada paper ini dilakukan analisis untuk melihat seberapa akurat segmentasi gambar dengan menggunakan algoritma random forest. Dalam makalah ini akan diulas tentang hasil perbandingan kinerja algoritma *Random forest* dengan algoritma J48, *Naïve bayes*, dan *Logistic regression*. Untuk hasil akurasi pada tiap-tiap algoritma tersebut dapat diukur berdasarkan pada recall, precision dan f1 score dari confusion matrix yang dihasilkan saat proses pengujian [8].

METODE

Algoritma Random Forest merupakan metode pembelajaran ansambel untuk klasifikasi, regresi, serta tugas pembelajaran mesin yang lain yang beroperasi dengan membangun banyak pohon keputusan pada waktu pelatihan. Untuk tugas klasifikasi yang termasuk dalam pembelajaran terawasi, hasil random forest dapat berupa kelas yang dipilih oleh sebagian besar pohon keputusan yang terbangun. Random Forest saat diaplikasikan mengalami overfitting, maka pohon keputusan akan kembali kepada set pelatihan mereka. Random forest pada umumnya mengungguli pohon keputusan tradisional seperti C45, namun akurasinya menjadi lebih rendah daripada pohon keputusan yang dioptimasi menggunakan gradien. Karakteristik data dapat mempengaruhi penampilan mereka [9][10].

Algoritma pertama untuk hutan keputusan acak dibuat pada tahun 1995 oleh Tin Kam Ho menggunakan metode subruang acak, dalam formulasi Ho terdapat cara untuk menerapkan pendekatan "diskriminasi stokastik" untuk klasifikasi yang diusulkan oleh Eugene Kleinberg.

Perpanjangan algoritma random forest dikembangkan oleh Leo Breiman dan Adele Cutler yang mendaftarkan "Random Forests" sebagai merek dagang pada tahun 2006 (per 2019, dimiliki oleh Minitab, Inc.)[11]. Hutan acak sering digunakan sebagai model kotak hitam dalam bisnis, karena menghasilkan prediksi yang masuk akal di berbagai data sementara hanya membutuhkan sedikit konfigurasi.



Gambar 1. Ilustrasi Algoritma Pengklasifikasi *Random Forest*

Pseudocode prediksi Random Forest dapat dijelaskan sebagai berikut:

Untuk melakukan prediksi menggunakan algoritma Random Forest yang diawali dengan kelompok data latih,

1. Mengambil fitur pengujian dan menggunakan aturan dari setiap pohon keputusan yang dibuat secara acak untuk memprediksi hasil dan menyimpan hasil yang diprediksi (target).
2. Hitung suara untuk setiap target yang diprediksi.
3. Pertimbangkan target prediksi pilihan tinggi sebagai prediksi akhir dari algoritma random forest.

Untuk melakukan prediksi menggunakan algoritma random forest terlatih, perlu melewati fitur uji melalui aturan dari setiap pohon yang dibuat secara acak [12]. Misalkan membentuk 100 pohon keputusan acak dari random forest. Setiap pohon acak akan memprediksi target (hasil) yang berbeda untuk fitur pengujian yang sama [13]. Kemudian dengan mempertimbangkan setiap target yang diprediksi, suara akan dihitung. Misalkan 100 pohon keputusan acak memprediksi sekitar 3 target unik x, y, z maka suara x tidak lain adalah dari 100 pohon keputusan acak berapa banyak prediksi pohon x. Begitu juga untuk 2 target lainnya (y, z). Jika x mendapatkan suara tertinggi. Katakanlah dari 100 pohon keputusan acak 60 pohon memprediksi targetnya adalah x. Kemudian hutan acak terakhir mengembalikan x sebagai target yang diprediksi. Konsep voting ini dikenal sebagai voting mayoritas [14].

Hyperparameters pada pengklasifikasi Random Forest dapat dijelaskan sebagai berikut;

1. `max_depth`: The `max_depth` dari pohon pada Random Forest didefinisikan sebagai jalur terpanjang antara simpul akar dan simpul daun.
2. `min_sample_split`: Parameter yang memberi tahu pohon keputusan di hutan acak jumlah pengamatan minimum yang diperlukan di node mana pun untuk membaginya. Default = 2
3. `max_leaf_nodes`: hyperparameter sets kondisi pemisahan node di pohon dan karenanya membatasi pertumbuhan pohon.
4. `min_samples_leaf`: Random Forest hyperparameter untuk menentukan jumlah minimum sampel yang harus ada di simpul daun setelah membelah simpul. Default = 1
5. `n_estimators`: jumlah pohon-pohon keputusan dalam 'the forest'.

6. **max_sample**: The **max_samples** hyperparameter menentukan bagian apa dari kumpulan data asli yang diberikan ke setiap pohon individu.
7. **max_features**: Ini menyerupai jumlah fitur maksimum yang disediakan untuk setiap pohon di random forest.
8. **bootstrap**: Metode pengambilan sampel titik data (dengan atau tanpa penggantian). Default = True
9. **criterion**: Fungsi untuk mengukur kualitas split. Kriteria yang didukung adalah "gini" untuk Gini impurity dan "entropy" untuk perolehan informasi.

Kelompok data gambar yang digunakan

Kelompok data yang digunakan pada paper ini adalah *Segmenting Challenge Dataset*, kelompok data ini telah diolah dan berasal dari aplikasi weka 3.6.8. Dataset ini merupakan data dari 7 gambar luar ruangan yang diambil secara acak untuk dibandingkan, apakah algoritma pengklasifikasi citra dapat menganalisis dengan akurat sesuai dengan jenis gambarnya (target/class-nya). Gambar-gambar itu disegmentasi untuk membuat klasifikasi untuk setiap pixel. Jumlah atribut pada kasus ini adalah 19 atribut yang berkelanjutan.

Memiliki 7 distribution classes yaitu brickface, sky, foliage, cement, window, path and grass. Distribution classnya membandingkan 7 kelas yaitu brickface (Tembok Bata), Sky (Langit), Foliage (Dedaunan), Cement (Semen), Window (Jendela), Path (Jalan), dan Grass (Rumput). Dari ketujuh kelas tersebut dipastikan bahwa tidak semua gambar akan menunjukkan secara sempurna maksud dari gambar tersebut.

HASIL DAN PEMBAHASAN

Hasil simulasi untuk nilai recall, precision dan F1 score kelompok data gambar

Berikut ini merupakan hasil simulasi 4 algoritma pengklasifikasi citra yang digunakan dalam penelitian ini. Berawal dari Confusion Matrix sebagai pengukuran kinerja untuk masalah klasifikasi pada pembelajaran mesin dimana keluaran dapat berupa dua kelas atau lebih. Confusion Matrix adalah tabel dengan 4 kombinasi berbeda dari nilai prediksi dan nilai aktual. Ada empat istilah yang merupakan representasi hasil proses klasifikasi pada confusion matrix yaitu True Positif, True Negatif, False Positif, dan False Negatif.

Hasil akurasi algoritma pengklasifikasi tersebut diukur menggunakan alat ukur kinerja antara lain *Recall*, *Precision*, dan *F1 Score*. Masing-masing alat ukur kinerja tersebut digunakan sesuai dengan fungsinya. *Recall* atau *Sensitivity (True Positive Rate)* *Recall* menggambarkan keberhasilan model dalam menemukan kembali sebuah informasi. *Precision* merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif. Skor F1 adalah nilai yang memberi tahu seberapa tepat algoritma pengklasifikasi (berapa banyak instance yang diklasifikasikan dengan benar), serta seberapa robust model (tidak melewatkan sejumlah besar instance).

Presisi tinggi tetapi *recall* yang lebih rendah, akan memberi hasil yang sangat akurat, tetapi kemudian dapat menimbulkan banyak *false positive* yang terjadi terutama jika recallnya sangat rendah. Persamaannya dapat dilihat pada persamaan ini. Sebaliknya jika *recall* tinggi tetapi presisi yang lebih rendah, akan melewatkan beberapa kejadian yang diinginkan tetapi lebih mengurangi kejadian *false positive*. Pada Machine Learning terutama di bidang Computer Vision, Metrics dapat diartikan sebagai sebuah nilai yang dapat digunakan untuk merepresentasikan performa model yang dihasilkan.

Tabel 1. Hasil perhitungan recall, precision dan F1 score algoritma *Logistic Regression*

%	A	B	C	D	E	F	G	HASIL	MEAN
Recall	97.1	100	91.4	96.6	90.1	100	100	96.4	
Precision	98.0	100	92.3	91.8	86.7	100	100	95.5	95.9

F1	97.5	100	91.8	94.1	88.1	100	100	95.9
----	------	-----	------	------	------	-----	-----	------

Tampak pada tabel diatas hasil pengukuran kinerja algoritma Logistic regression berdasarkan confusion matrix yang dihasilkan. Akan dilakukan perbandingan nilai recall, precision dan skor F1 dengan algoritma yang lain.

Tabel 2. Hasil perhitungan recall, precision dan F1 score algoritma *Naïve Bayes*

%	A	B	C	D	E	F	G	HASIL	MEAN
Recall	75.7	100	73.5	84.4	47.2	98.2	99.0	82.5	80.4
Precision	90.2	99.0	18.7	86.3	72.0	94.4	100	80.0	
F1	84.1	99.5	29.8	85.3	57.0	96.3	99.5	78.7	

Pada tabel diatas hasil pengukuran kinerja algoritma Naïve Bayes berdasarkan confusion matrix yang dihasilkan. Akan dilakukan perbandingan nilai recall, precision dan skor F1 dengan algoritma yang lain.

Tabel 3. Hasil perhitungan recall, precision dan F1 score algoritma J48

%	A	B	C	D	E	F	G	HASIL	MEAN
Recall	97.5	99.5	89.4	95.4	89.0	99.5	100	95.7	95.1
Precision	95.6	100	94.2	94.0	87.7	98.7	99.0	95.6	
F1	96.5	99.7	91.8	94.3	88.3	88.3	99.5	94.0	

Berikutnya tabel diatas merupakan hasil pengukuran kinerja algoritma J48 berdasarkan confusion matrix yang dihasilkan. Akan dilakukan perbandingan nilai recall, precision dan skor F1 dengan algoritma yang lain.

Tabel 4. Hasil perhitungan recall, precision dan F1 score algoritma *Random forest*

%	A	B	C	D	E	F	G	HASIL	MEAN
Recall	99.5	99.5	94.8	96.3	94.9	99.5	100	97.7	97.7
Precision	99.0	100	96.6	96.8	92.6	99.5	100	97.7	
F1	99.2	99.7	95.7	96.5	93.7	99.5	100	97.7	

Terakhir pada tabel diatas merupakan hasil pengukuran kinerja algoritma Random Forest berdasarkan confusion matrix yang dihasilkan. Akan dilakukan perbandingan nilai recall, precision dan skor F1 dengan algoritma yang lain.

KESIMPULAN

Dari simulasi yang dilakukan menghasilkan nilai perbandingan antara logistic regression dan naïve bayes masing- masing algoritma memiliki keakuratan sebesar 95.9% dan 80.4%. Dapat dilihat bahwa selisih antara keduanya hingga 15.5% yang berarti bahwa logistic regression memiliki kinerja 15.5% lebih baik daripada menggunakan naïve bayes untuk menganalisis gambar sesuai jenisnya. Sedangkan perbandingan antara J48 dan random forest masing-masing memiliki keakuratan masing-masing sebesar 95.1% dan 97.7%. Dapat dilihat bahwa selisih antara keduanya mencapai 2.6% yang berarti bahwa penggunaan kedua algoritma untuk menentukan keakuratan sebuah gambar berdasarkan klasifikasinya tidak berbeda jauh dan kinerja 2.6% lebih baik dengan menggunakan random forest. Serta hasil perbandingan sebelumnya didapatkan bahwa dari ketujuh kelas terdapat 2 sub kelas yang sulit untuk dibedakan oleh algoritma-algoritma yang digunakan dalam simulasi yaitu brickface dengan cement dan foliage

dengan window. Pada perbandingan klasifikasinya diketahui bahwa hasil klasifikasi dengan keakuratan terendah pada algoritma Naïve bayes dan nilai tertinggi pada algoritma Random forest.

DAFTAR PUSTAKA

- [1] Z. Li *et al.*, “Pixel-Level Segmentation of Bladder Tumors on MR Images Using a Random Forest Classifier,” *Technol. Cancer Res. Treat.*, vol. 21, no. 169, pp. 1–9, 2022, doi: 10.1177/15330338221086395.
- [2] A. Subudhi, M. Dash, and S. Sabut, “Automated segmentation and classification of brain stroke using expectation-maximization and random forest classifier,” *Biocybern. Biomed. Eng.*, vol. 40, no. 1, pp. 277–289, 2020, doi: 10.1016/j.bbe.2019.04.004.
- [3] T. Yang, J. Song, and L. Li, “A deep learning model integrating SK-TPCNN and random forests for brain tumor segmentation in MRI,” *Biocybern. Biomed. Eng.*, vol. 39, no. 3, pp. 613–623, 2019, doi: 10.1016/j.bbe.2019.06.003.
- [4] Y. Wu and S. Misra, “Intelligent Image Segmentation for Organic-Rich Shales Using Random Forest, Wavelet Transform, and Hessian Matrix,” *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 7, pp. 1144–1147, 2020, doi: 10.1109/LGRS.2019.2943849.
- [5] V. R. Balaji, S. T. Suganthi, R. Rajadevi, V. Krishna Kumar, B. Saravana Balaji, and S. Pandiyan, “Skin disease detection and segmentation using dynamic graph cut algorithm and classification through Naive Bayes classifier,” *Meas. J. Int. Meas. Confed.*, vol. 163, p. 107922, 2020, doi: 10.1016/j.measurement.2020.107922.
- [6] Z. Zhang and Y. Han, “Detection of Ovarian Tumors in Obstetric Ultrasound Imaging Using Logistic Regression Classifier with an Advanced Machine Learning Approach,” *IEEE Access*, vol. 8, pp. 44999–45008, 2020, doi: 10.1109/ACCESS.2020.2977962.
- [7] N. N. et al. . Nora Naik et al., “Detection and Classification of Brain Tumor Using Naïve Bayes and J48,” *Int. J. Comput. Sci. Eng. Inf. Technol. Res.*, vol. 9, no. 2, pp. 19–28, 2019, doi: 10.24247/ijcseitrdec20194.
- [8] A. Murugan, S. A. H. Nair, and K. P. S. Kumar, “Detection of Skin Cancer Using SVM, Random Forest and kNN Classifiers,” *J. Med. Syst.*, vol. 43, no. 8, 2019, doi: 10.1007/s10916-019-1400-8.
- [9] T. K. Ho, “Random decision forests,” *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, vol. 1, pp. 278–282, 1995, doi: 10.1109/ICDAR.1995.598994.
- [10] T. Ho, Kam, “The Random Subspace Method for Constructing Decision Forests,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, 1998, [Online]. Available: %3CGo%0Ato
- [11] W. Deng, Z. Huang, J. Zhang, and J. Xu, “A Data Mining Based System for Transaction Fraud Detection,” *2021 IEEE Int. Conf. Consum. Electron. Comput. Eng. ICCECE 2021*, pp. 542–545, 2021, doi: 10.1109/ICCECE51280.2021.9342376.
- [12] P. Geurts, D. Ernst, and L. Wehenkel, “Extremely randomized trees,” *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, 2006, doi: 10.1007/s10994-006-6226-1.
- [13] T. K. Ho, “A data complexity analysis of comparative advantages of decision forest constructors,” *Pattern Anal. Appl.*, vol. 5, no. 2, pp. 102–112, 2002, doi: 10.1007/s100440200009.
- [14] S. Wintner, “Dietterich TG: An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees,” *En.Scientificcommons.Org*, pp. 139–157, 2000, [Online]. Available: <http://en.scientificcommons.org/42637098%5Cnuuid/7906280C-AEF8-405A-9A94-6BAA1DDAED1E>